THE LISTER HILL NATIONAL CENTER
FOR BIOMEDICAL COMMUNICATIONS

An Intramural Research Division of the U.S. National Library of Medicine

A Report to the Board of Scientific Counselors
September 2013

Clinical Text De-Identification Research

Mehmet Kayaalp, MD, PhD, Lead Investigator, CgSB
Allen Browne, MS
Zeyno A. Dodd, PhD
Pamela Sagan, BSN, RN
Clement J McDonald, MD

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

Table of Contents

Introduction

The Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA) requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. We have been developing a software tool to de-identify clinical records, which we have named NLM Scrubber. Version 1.0 of the system currently recognizes and redacts patient names, alphanumeric identifiers, addresses and dates. NLM Scrubber's success rate of de-identifying these identifiers is around 99% and its rate of conserving text of health information with no personal identifiers is 99%, without counting de-identified provider names as false positives. We plan to release the system as an open source tool in early 2014.

Background

Electronic health records are treasure troves for clinical scientists because with the availability of high volumes of electronic reports, clinicians are no longer limited to a cohort of their patients and can easily test their hypotheses on much larger samples. Access to those records, however, is not easy and involves overcoming a number of institutional barriers. These barriers have been raised purposefully to ensure that only the right person would access private information of the patient. Access is warranted only when necessary justifications for the study and other assurances are provided that the proposed study is scientifically sound and important for the greater patient population, and the protocol is safe, secure and well planned.

While these barriers had been the primary tool to protect patient privacy, the requirements were so difficult to attain that they become a barrier before the scientific progress. Having seen both sides of the issue, in 1991 the U.S. Congress enacted HIPAA where it has tasked the U.S. Department of Health and Human Services (HHS) to regulate access to health records while protecting the health information of individuals.

HIPAA Privacy Rule

As defined by HHS, Protected Health Information (PHI) comprises a subset of health information of an individual who is the subject of the health record *and* the information is associated personally identifiable information[1] (PII), including demographic information, collected from the individual to be used by the health care provider, health plan, employer or health clearinghouse. PII is any information that distinguishes or traces an individual's identity such as name, social security number, date of birth or biometric records and any other information such as medical, financial and employment information that is linkable to an individual.[2 3] HHS developed the Privacy Rule, where it defined certain identifiers as part of PHI, which should be de-identified before health records are accessed for research purposes (see Table 1). Note that the health information dissociated from those identifiers of the individual is not considered PHI. According to the Privacy Rule the identifiers in Table 1 that belong to the individual or relatives, employers or household members of the individual, should not be present in any de-identified health records.[4]

Table 1 Per HIPAA Privacy Rule, the following identifiers must be deleted from PHI to fully de-identify health information. (*) As of 2010, there were 18 sets of zip codes with distinct initial three digits whose corresponding population sizes were less than or equal to 20,000.[1]

1. Names
2. All geographic subdivisions smaller than a state, except the first two digits of the zip code of the postal address. The third digit of the zip code can also be left intact, only if the size of the population in the area of the censored two digits is greater than 20,000 according to the most recent census data.(*)
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Telephone numbers.
5. Fax numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan beneficiary numbers.
10. Account numbers.
11. Certificate/license numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet Protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, except the ones that may be generated by the covered entity for re-identification.

[1] The text of CFR 45 § 164.514 uses the term *individually* identifiable information instead of *personally* identifiable information. One possible reason is that the meaning of the legal term *person* also includes entities other than natural person (human) such as trust, estate, partnership, corporation, and professional association among others. On the other hand, personally identifiable information and its acronym PII are more widely known terms; hence, they are used in this report instead.

Limited Data Set

The Privacy Rule applies only to covered entities, which are health plans, health care clearinghouses, or health care providers who transmit any health information in electronic form.[2] A covered entity may use or disclose a *limited data set* without the written authorization of the individual for the purposes of research, public health, or public health operations. A limited data set may not contain any of the identifiers in Table 1, except town or city, state, and zip code (as part of the postal address) and dates (e.g., dates of birth and death, dates of health care services, including hospital admission and discharge as well as individual's age in year, month, day, and time).[4] In other words, unlike fully de-identified data set, a limited data set may also contain the following PII: all dates and ages as well as the full zip code and town information of the address. Given the presence of certain dates and/or postal address information (except street address) related to the individual, a limited data set is PHI, and the recipient of the limited data set has to sign into a data use agreement with the covered entity. The requirements of a data use agreement are specified in 45 CFR § 164.514(e)(4).[4]

Current Text De-identification Systems

De-identification of a structured data is a fairly straightforward process, where fields containing PHI should be identified and their contents should be deleted or made inaccessible to researchers. De-identification of an unstructured data or free text, on the other hand, is a rather challenging task. Because of the idiosyncrasies of any natural language, including English, the utterances of information are not always predictable and we have to devise intelligent tools to recognize those words and phrases containing PHI.

A thorough review of 18 clinical text de-identification systems has been published recently.[5] Since then only two other new systems appeared in major journals.[6 7] These 20 systems can be categorized in two groups based on their target documents: general purpose vs. niche (specialized) de- identifiers. They can also be classified in terms of their underlying methodologies, which roughly are symbolic or machine learning approaches. Symbolic approaches mainly rely on rules, regular expressions, and lookup tables (also referred to as dictionaries or gazetteers). The availability of a de-identification system is another important characteristic; some are freely available, some are commercial products, and others have not been made available.

Currently, there are only five freely available systems, three of which were specialized to de-identify surgical pathology reports only.[8-10] The other two systems are general purpose de-identification systems developed by researchers at MIT and MITRE. MIT's system took a symbolic approach; whereas, MITRE's is a machine learning system using conditional random fields.

The name of the MIT's system was not mentioned in their publication[11] but the filename of the code was *deid.pl*. Since there is another (commercial) system with the same name, De-ID, to prevent any confusion, we here call MIT's system MITdeid. MITdeid provides various features that are closely tuned to clinical setting, such as accepting a list of provider names of the institute and the full name of the patient per report.

The MITRE's system, MIST, was developed to demonstrate how an existing conditional random field program designed for a generic use could be repurposed quickly as a successful clinical text de-identification system.[12] MIST has proven itself as one of the most successful systems in the i2b2 competition in 2006.[13] As a machine learning system, MIST requires a training dataset. The current version of the system does not store the constructed model and has to be re-trained before each testing session.

NLM Name Scrubbing Study

As part of this project, we studied personal name recognition in great depth.[14] In this study, we analyzed dictated clinical notes and imaging study reports with the focus on personal names, namely patient and provider names. We considered not only actual patient names but also the names of the relatives, the household members, and employers of the patient as the patient name. We studied the prevalence of these personal names in various report types and how well our system along with other de-identification systems (MIST and MITdeid) and prominent NLP tools (LingPipe and ANNIE) perform in recognizing personal names in clinical text. We also studied the performance of these systems in three modes: (1) with no extra information (outside the clinical text report) provided to the systems, (2) available patient names and provider name roster provided, and (3) in addition to the patient and provider names, NLM name datasets provided.

Project Objectives

The objective of this project is to build a clinical text de-identification system. Our broader goal is to promote scientific progress in biomedicine by enabling researchers to access large amount of de-identified health information. While we have focused on the development of a stand-alone application for de-identification, we are also considering alternative approaches such as de-identification as an online service. We can also consider collaborations with clinical institutions to help them create large collections of de-identified health information to be used by a wider research community.

Project Significance

There have been several attempts to de-identify clinical text data automatically via software, but none of the freely available tools is good enough to lower the risk of privacy to an acceptable minimum level. As part of HHS, NLM started the clinical text de-identification project to respond to this need and promote scientific progress by enabling the research community to access large amount health information that do not contain personal identifiers.

Significance of our project depends on the degree to which we can facilitate the production of de-identified clinical text data, and minimize (if not eliminate) the burden of manual de-identification for the clinical research community.

Methods and Procedures

In this section, we present (1) how we select and process the clinical text data from a large corpus of clinical reports in order to reliably study and develop robust de-identification methods; (2) the methods and components of our de-identification tool; and (3) our evaluation methods.

Data Selection

A typical hospital information system preserves every copy and version of a clinical report, yielding a large number of duplicate narrative texts. A sample with duplicate reports may inflate the magnitude of events observed in the study. If the duplication is randomly distributed, it would increase the noise, but if it is biased in a particular direction, the results would be erroneous and misleading. To ensure the reliability and the robustness of the study results, we devised a random sampling method to exclude both fully and partially duplicate reports from the study data—we classify two distinct versions of the same report as partially duplicate.

The basic premise of the method is to limit the inclusion criterion to one report per report type per patient visit with the latest timestamp. For each randomly selected patient, we collected all reports generated during a particular visit of the patient, clustered them by report types, sorted each cluster by report filing time, and took only the latest, presumably the most developed report in that cluster.

Our sampling method relies on the assumption that each visit is associated with a unique visit number and reports of the same type in two different visits are sufficiently dissimilar. Note that this assumption may not always hold. After performing the sampling, we sorted reports of each patient by word counts. The manual comparison of the reports that are similar in size helped us discover two sets of reports, where two reports in each set with almost identical content were associated with different visits of the same patient. We eliminated the earlier reports from these sets. This sampling method may inadvertently eliminate some non-duplicate reports, but in the final analysis, it yields an unbiased, large spectrum of reports per visit with distinct report types.

In the study of personal name de-identification, we used 3093 distinct clinical reports about 1636 distinct patients of the Clinical Center at NIH. The maximum number of reports per patient was 20. The distribution of the report types in the study data is listed in Table 2.

We developed NLM Scrubber using 1140 clinical reports from the same origin, which we call the training data. Unlike the study test data, retrieval of the training data was done in several iterations over a long period of time in an ad hoc and not truly randomized manner.

Annotation

In order to evaluate our de-identification methods, we needed to create a set of manually annotated reports to be used as our gold standard. Toward this end, we developed a PII annotating application, called Visual Tagging Tool (VTT).[15] VTT was designed in conjunction with this project has been released to the community at large as one of the SPECIALIST NLP Tools. VTT takes plain text as its input. Using the graphical user interface (GUI) and the mouse, the annotator can select any contiguous portion of text and choose a markup tag from a menu of items to tag the selected portion of the text. The annotator can also change the tag set using GUI

Table 2 Decomposition of the Clinical Narrative Reports in the Study Data

Physician Observation Reports	Count	Ratio
Discharge Summary	270	25%
Consultation Report	245	23%
First Registration Report	123	11%
History & Physical Exam	92	9%
Discharge Summary w/ HPE	82	8%
PT Results	57	5%
Outpatient Single Visit	50	5%
OT Results	48	4%
Inpatient Operation	26	2%

Patient Study Reports	Count	Ratio
DX	614	30%
CT	449	22%
MRI	214	11%
US	182	9%
DEXA	173	9%
PET	138	7%
NM	127	6%
MM	24	1%
SP	22	1%
EEG	18	1%

Consult-Final Only	17	2%
Outpatient Operation	16	1%
Rehab Medicine Results	11	1%
Outpatient Addendum	11	1%
Radiation Oncology HPE	10	1%
Interim Summary	4	0%
Outpatient Summary	4	0%
Rad. Therapy Summary	3	0%
Addendum Summary	2	0%
Interim Summary w/HPE	2	0%
Total	**1073**	**100%**

EMG	16	1%
MRA	15	1%
IP	10	0%
PETR	9	0%
Holter	5	0%
FL	4	0%
Total	**2020**	**100%**

directly. Each tag is associated with a distinct visual pattern that is also customizable. For example, in Figure 1, the following patterns are displayed:

- **untagged text**: text left in its original format
- **alphanumeric identifier**: italicized characters on pink background
- **date identifier**: underlined characters on yellow background
- **personal name identifier**: characters in bold white font on red background
- **age identifier**: underlined characters in white font on dark green background
- **organization identifier**: characters in bold font on bright blue background
- **de-identified text containing no PHI**: characters in bold fonts on grey background

Unlike many other annotation tools, VTT conserves the original text as is. The specification of each tag type is

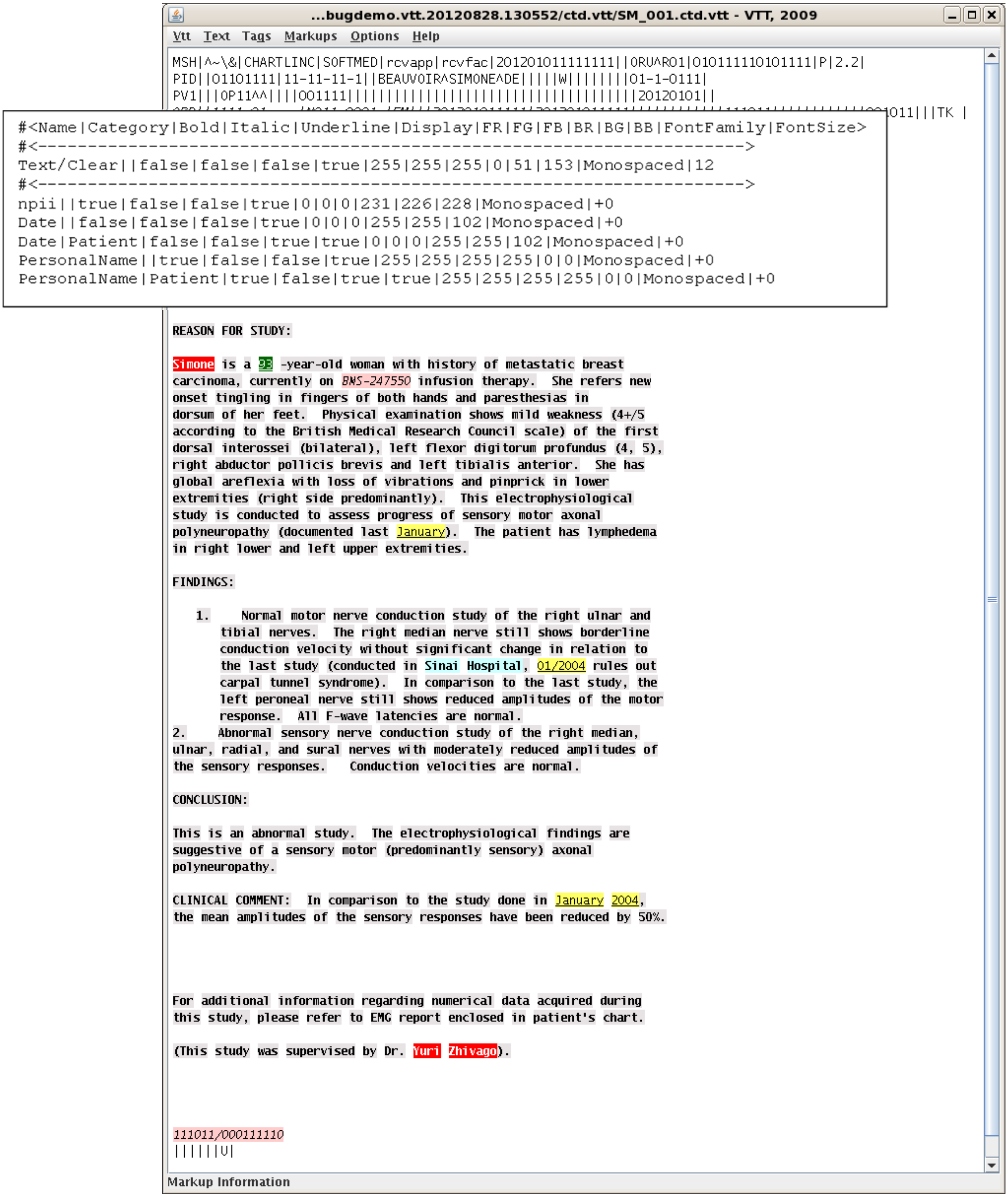


Figure 1 Annotated Fictitious HL7 Message Displayed Using VTT

appended to the end of the original text in a single line. The specification includes the name of the tag, a subcategory (e.g., patient) if applicable, the font characteristics, RGB combinations for font and background colors (see Figure 2).

The tag specifications are followed by the annotation section. Annotations are ordered as they appear in the original text. Each annotation line contains the information of a single tagged text, including offset, length of the tagged character string, the tagged text (token or phrase), and associated annotations (see Figure 3).

```
499| 9|AlphaNumericId |Patient|Z011-0001
518| 8|npii           |       |Dictated
527| 2|npii           |       |by
529| 1|npii           |       |:
535| 7|PersonalName   |       |Gregory
543| 1|PersonalName   |       |A
545| 5|PersonalName   |       |House
550| 1|PersonalName   |       |,
552| 4|npii           |       |M.D.
581| 4|npii           |       |Exam
586| 4|npii           |       |Date
590| 1|npii           |       |:
593|10|Date           |Patient|01/01/2012
613| 6|npii           |       |REASON
620| 3|npii           |       |FOR
624| 5|npii           |       |STUDY
629| 1|npii           |       |:
640| 6|PersonalName   |Patient|Simone
647| 2|npii           |       |is
650| 1|npii           |       |a
652| 2|Age-PII        |Patient|93
655| 1|npii           |       |-
656| 4|npii           |       |year
660| 1|npii           |       |-
661| 3|npii           |       |old
665| 5|npii           |       |woman
```

Figure 3 Sample of Annotated Text

The VTT format was designed so that the resulting text is easily readable without the VTT GUI, easily traceable using other codes, and easily reproducible by others without requiring major coding effort. Two different de-identification modules (e.g., name and address de-identifiers) can independently tag the same token with different tags. The conservation of the original data, the separation of the data from its interpretations,[2] and accepting multiple (sometimes conflicting) interpretations of the data were our annotation format requirements based on the design principles that we outlined in a technical report[16] prior of the inception of this project.

Supporting datasets

We obtained from the Office of the Chief Actuary (OCA) of the U.S. Social Security Administration two large collections of personal names containing 1,096,440 unique first and 2,192,183 unique last names associated with population frequencies. The data were derived from 448 million people who applied for social security

[2] Note the tags/annotations are particular interpretations of the data by a de-identification system or an annotator.

numbers. Both first and last name datasets exclude names whose population frequencies are less than 3. These two sets were disassociated; that is, full names (i.e., first and last name pairs) were not made available.

Our second personal name dataset is known as the Social Security Administration's Death Master File (DMF).[17] DMF contains PII of the deceased U.S. population collected since 1936. The DMF population was the subset of the population of the OCA data. Unlike the OCA data, DMF was uncensored. Our copy of DMF received in 2008 contains the full names of 80,579,812 individuals.

Our third personal name dataset was extracted from the author field of the MEDLINE® dataset.[18] Since population frequencies of these names were not available, we made a heuristic assumption that they are located in the censored portion of the OCA data. We heuristically assigned each author name that is unobserved in other name sets with a frequency count of 1, if seen only once in MEDLINE, or 2, otherwise.

Since our algorithm does not make any distinction between first and last names, we combined all personal name sets into one dataset containing 3,833,957 unique personal names.

The two English corpora used in this study were Wikipedia (English)[19] and the abstracts of MEDLINE articles from 121 core clinical journals (CCJ).[20] Both corpora, which we compiled in 2008, contain over 1 billion tokens each. Wikipedia and CCJ contain over 5 million and 2 million unique tokens, respectively.

De-identification Methods

In this section, we present the de-identification methods of NLM Scrubber by its components, which are patient name, alphanumeric ID, date/age, and address recognition and redaction of the PHI tags from text. A large part of our effort was devoted to personal name recognition.

Patient name recognition

Dictionary-based approaches usually fall short in recognizing names in text, because many names are spelled like regular English words. To address this problem, we used the likelihood ratio metric to detect likely personal names along with the initCap rule, the requirement that tokens can be labeled personal names only if their initial letters are capitalized. The only exception to initCap rule is the set of tokens that are known as nobiliary particles such as **von**, **van**, **de**, **di**, **dos** etc. Similar to the initCap rule, we also used the *noDigit* rule, which filters out words containing digits. In this study, we also used a simple filtering system to retain a limited set of high frequency clinical words and to reduce the total false positive count.

Since the prerequisite statistics for calculating the likelihood ratio metric were not available, we devised a new method to estimate the components of the metric from statistics of two sets of distinct samples.

Likelihood ratio metric

We introduced a new way of estimating likelihood ratio statistic that helped us to de-identify 98% of the patient names correctly. The concept of likelihood ratio originates from Bayesian Theory—$posterior\ odds = prior\ odds\ \times likelihood\ ratio$.[21] Likelihood denotes the probability of an observation within a given context. In clinical decision making, likelihood ratio (1) is formulated usually as the ratio of probabilities of observation D, given two competing hypotheses H_1 and H_2, which may refer to a disease and no-disease, respectively.

$$\frac{P(D|H_1)}{P(D|H_2)} \qquad (1)$$

In clinical diagnosis, D simply refers to a clinical symptom and/or sign. In personal name recognition, H_1 states that a particular token/word D is a name; whereas, H_2 is its complement stating the opposite (i.e., D is *not* a name). Conventionally, both the numerator and denominator terms of a likelihood ratio are estimated from the same dataset. For example, given a cohort of patients with a certain radiologic finding, some patients may develop cancer but others do not. The ratio of these patients would be the likelihood ratio.

Unlike in clinical context, we do not have prior studies reporting such statistics for our domain. We attempted to estimate these probabilities based on the following simple observations.

1. Common names such as JOHN belong to a very large portion of the population; thus, observing such common names in a regular corpus of text is proportionally more likely than observing less common names in the same corpus. Although this proportionality does not hold for celebrities, historical figures, and well-known fictional characters, we expect that it would hold for the majority of names.
2. Based on personal name frequencies of social security applicants, we can estimate the probability of a particular name among all names in our dataset. So we estimate the numerator's likelihood in (1).
3. Using a corpus of clinical narratives, we can estimate the probability of a word among all words in the corpus. Since we did not have a large spare corpus of clinical narratives that was mutually exclusive from the corpus of our experiments, we used Wikipedia (English edition) and Medline abstracts of core clinical journals as our corpus. Given that personal names are relatively rare in our corpus, we made yet another simplification assumption that they would not significantly alter the likelihood of regular words. So we estimate the denominator's likelihood in (1).

Given that necessary statistics do not exist in our domain, we had to introduce this new method to estimate likelihood ratios from disparate samples.

We estimated probabilities in (1) through Bayes-Laplace smoothing (2), where N denotes sample size, $n(D)$ frequency count of event D in the sample, and $1/r$ is a Bayes-Laplace prior:[22]

$$P(D) = \frac{n(D) + 1}{N + r} \tag{2}$$

In the likelihood ratio, probabilities of D are conditional to H_1 and H_2. The term $P(D)$ in (2) becomes conditional as we parameterize the equation using the corresponding (personal names vs. English corpus) dataset. Because of the smoothing function, resulting probabilities are always greater than 0 even if D does not occur in the sample.

Also note that for various instances of D, the denominator in (2) does not change for either sample. Since the ratio of the denominators stays constant for all estimates, the (increasing or decreasing) order of likelihood ratios of a sequence of words does not change for different samples (e.g., between two corpora in distinct genres or medical reports) as long as the relative frequencies of words in those samples (e.g., $n(JOHN) > n(MARY)$) stay proportional.

$$\frac{P(D|H_1)}{P(D|H_2)} \propto \frac{n(D|H_1)+1}{n(D|H_2)+1} \quad (3)$$

By assuming the same or similar orders of names frequencies both in clinical reports and English text corpora, we could use Wikipedia and Medline abstracts as proxies for a large clinical corpus.

The likelihood ratio behaves as follows:

1. If D is a name and does not refer to another sense in a clinical document, the likelihood ratio would yield a score much larger than 1.
2. If D is ambiguous, the likelihood ratio still yields a score greater than 1 for most of the names.
3. If D is not a name or it is an infrequent name and as a regular word it occurs disproportionally more frequently in the corpus, the likelihood ratio would yield a score lower than 1.
4. If D is found neither in the name dataset nor in the corpus (i.e., the right-hand-side of proportionality (3) equals 1), due to the sample sizes and our choice of Bayes-Laplace priors, the likelihood ratio, which at that point equals $(N_2 + r_2)/(N_1 + r_1)$, would yield a score of 2.4, favoring for personal name hypothesis.

Multi-token name recognition using context

Note that like *initCap* and *noDigit* rules, likelihood ratio metric is a token-centric approach, and does not consider context. For those tokens (e.g., **May**) that do not look like personal names to the likelihood ratio metric, we need to check the surrounding tokens to understand the context in which the token was used. Such a method would enable us to catch names (e.g., **May Smith**) that were not labeled as names by the likelihood ratio metric, but were colocated with other labeled names.

Based on this insight, we devised a method based on automata theory.[23] We developed a finite state automaton (FSA) representing a simple personal name pattern (see Figure 4). It is composed of six states: start (**S**), prefix title (**P**), initial (**I**), name (**N**), suffix title (**X**), and end (**E**). Each state can take one of the following six inputs: prefix (**p**), initial (**i**), name (**n**), suffix (**s**), punctuation (**!**), and anything else ($*$). If an input is a single uppercase letter, it is labeled as initial (i.e., **i**). We have four other mutually exclusive sets (tables) for prefixes, suffixes, names and punctuation marks. If an input token is a member of one of these sets, it is labeled accordingly (i.e., **p**, **s**, **n**, and **!**, respectively); otherwise, it is labeled as $*$.

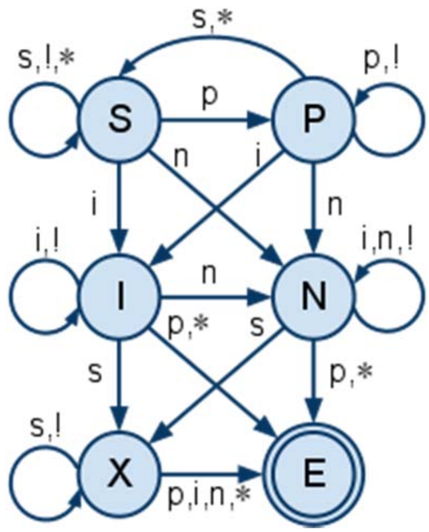


Figure 4 Finite State Automaton for Personal Name Recognition

If a sequence of tokens is accepted by FSA, tokens that correspond to name states are labeled as personal names. For example, if the input tokens were "**Mr . John A . Smith came**" (i.e., <**p**,**!**,**n**,**i**,**!**,**n**,∗>) and FSA was initially in state **S**, state transition sequence would have been <**S**,**P**,**P**,**N**,**N**,**N**,**E**>. An input sequence is identified as a name by FSA if and when the state transition reaches state E.

Alphanumeric identifier recognition

We define an alphanumeric string as a string of characters containing at least one or more digits. It may or may not contain other characters. Alphanumeric Identifier Recognizer (AIR) has a two-prong approach: It detects patterns that correspond to alphanumeric strings such as phone and social security numbers that need to be labeled as such, but it also detects patterns of known clinical entities that need to be preserved. AIR also attempts to distinguish alphanumeric strings from date-like patterns so that dates would not be mislabeled.

AIR ignores tokens that do not contain two or more digits; otherwise, it analyzes the content of token t and its context. If t is preceded by a token containing certain strings such as number, protocol, or #, it labels t as an alphanumeric identifier. If an alphanumeric string containing a sequence of two or more upper case letters, followed by certain tokens such as protocol, it is labeled as alphanumeric identifier. A 9–10 digit number patterns with or without delimiters in between are detected as alphanumeric identifiers (i.e., phone or social security numbers).

AIR also checks numerous conditions (e.g., a number followed by a unit of measure) that may indicate that token t is a valid piece of clinical data and should be conserved. AIR marks most other alphanumeric strings as alphanumeric identifiers.

Date and age recognition

Algorithms for identifying dates and ages are based on a set of regular expressions to detect the corresponding patterns. Some date patterns are listed in Table 3. For example, string 07-08-2012 would be identified using the pattern **DD*MM*YYYY**, where *s are delimiters and **D**, **M**, **Y** are digits such that YYYY should be greater than 1900 and less than the current year, $1 \leq \text{DD} \leq 31$ and $1 \leq \text{MM} \leq 12$.

Table 3 Date Patterns. D, M, Y, h, and m are date, month, year, hour and minute digits; * is a delimiter; MONTH, HOLIDAY are literal values of month and holiday incl. abbreviations; X? denotes that X is optional; | concatenates choices

Pattern	Example
YYYY*MM*DD	2012-08-07
DD*MM*YYYY	07-08-2012
MM*DD*YYYY	08-07-2012
MM*DD*YY	08-07-12
M*DD*YY	8-07-12
YYYY*YYYY	2011-2012
DD*MM*YY	07-08-12
M*D*YY	8-7-12
M*DD*YY	8-07-12
MM*YYYY	08-2012
DD*MM*DD*MM	07-08-08-08,07-08/08-08
MM?*DD	08-07, 8-07
DD?*MM	07-08, 7-08
MM*DD?	08-07, 08-8
DD*MM?	07-08, 07-8
YYYYMMDD	20120708
YYYYMMDDhhmm	201207081215
YYYY	2012
DD?*?MONTH	7-August, 7August, 7 Aug
MONTH*YY(YY)?	August.2012, August'12, Aug-12
(early|mid|late)*YYYY	Mid-2012
YYYY*?MONTH	2012/August, 2012Aug
'YY*?MONTH	'12-August, '12Aug
DD?MONTH*YY	7August'12
MONTH*DD?	Aug7, August 7
MONTH	Aug, August
HOLIDAY	Christmas, Easter

Unlike date patterns, age patterns are more involved. For example, age patterns may require to catch phrases like "on his **ninety-third** birthday" or "in his late **90ies**".We classified alphanumeric age expressions and label them with specific names (see Table 4). The corresponding patterns are recognized through regular expressions.

Table 4 Alphanumeric Age Expression Classes

Expression Classes	Examples
AGE-WITH-SUCCEEDING-MARKER	he was [93 years-old]
AGE-WITH-PRECEDING-MARKER	at the [age of 93],
AGE-WITH-APPENDED-UOM	his father, [93yo], has
AGE-FRACTION-EXPR	he is [5-years, and 3-months] old
AGE-FROM-PHRASE-CONTEXT	she [was nearly 93].
AGE-BIRTHDAY-CONTEXT	on his [ninety-third birthday]
AGE-DECADE-CONTEXT	in his late [90ies]

AGE-SIMPLE-CATCH-ALL	(as 93)
AGE-COMPOUND-CATCH-ALL	(93 and 90)

Whenever a date (age) regular expression is matched with the tokens in the text, those tokens are labeled as date (age).

Address identifier recognition

Addresses are recognized mostly via shapes of dTagger, a specialized part-of-speech tagger extended with limited pattern tagging abilities for entities, such as addresses. The dTagger searches address terms in various lexicons, which contain city and states names as well as street types and their abbreviations (e.g., Avenue, Alley, Blvd, and Circle). In its current format, the recognizer is difficult to maintain and will be revised before the release of the software package; therefore, we do not provide any further specifications of the soon-outgoing recognizer in this report.

Redaction

Redacting is a post-processing step, whose input is a tagged text. It outputs the de-identified text where tagged text content is replaced with the corresponding tag labels (see Figure 5). If two distinct recognizers tag the same token, the redactor labels the content as PHI instead of choosing one tag over another.

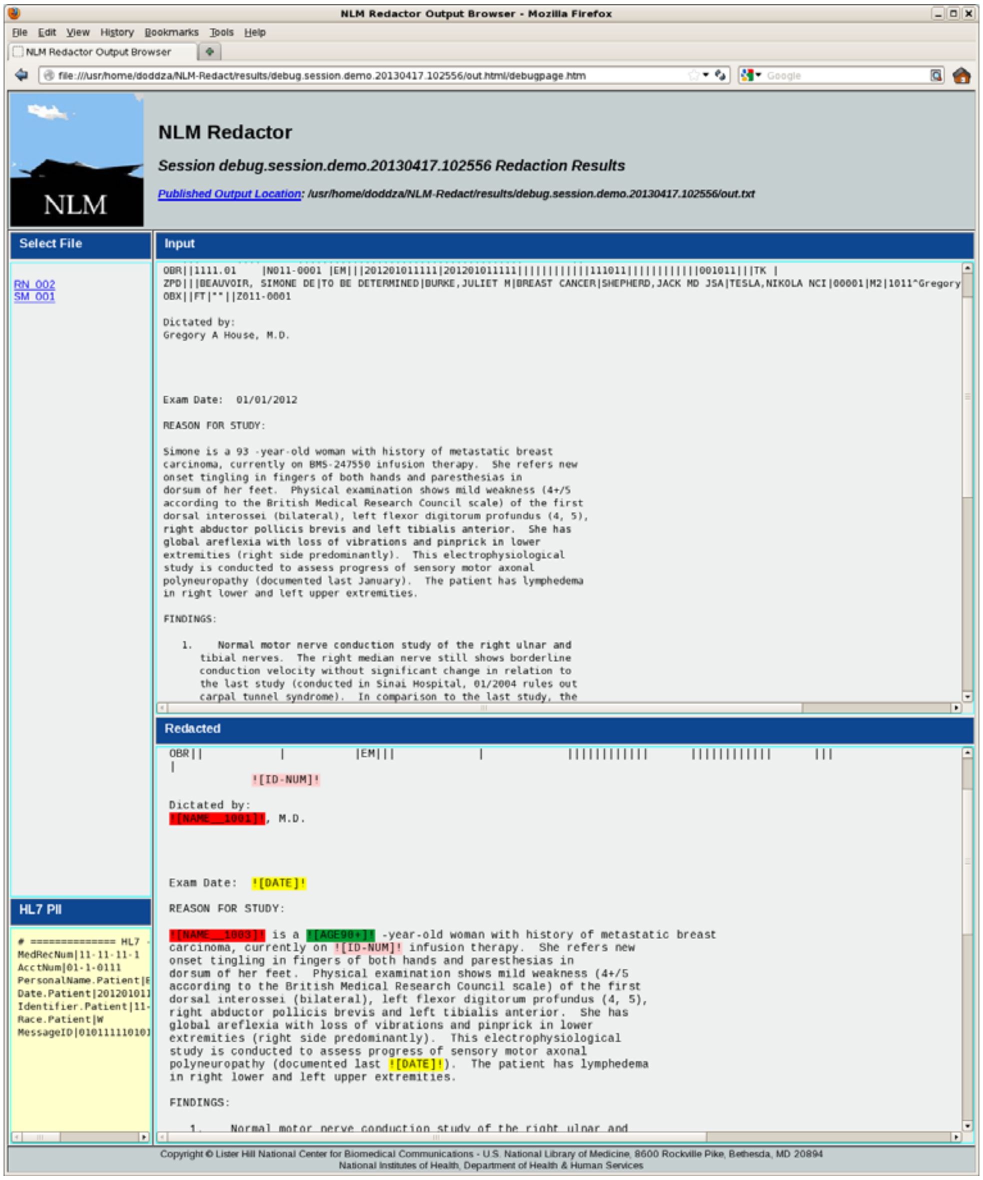


Figure 5 Example of Input and Output of PHI Redactor

Evaluation Methods

We evaluated the NLM Scrubber on a test set of 3093 dictated narrative reports generated at the NIH Clinical Center. The set was annotated by two experts, a linguist and a registered nurse, producing the gold standard for the test data. Following NLM Scrubber's run on the test data, we compared the resulting tags against the gold standard and evaluated them in terms of sensitivity, specificity and accuracy. We also evaluated the privacy risks due to the revealed PHI tokens.

Two most prominent and freely available de-identification systems, MIST and MITdeid, were tested on the same data. Their results were evaluated in terms of sensitivity, specificity and accuracy as well.

Since MIST is a machine learning system, it requires training before testing. We used our held-out set of 1140 annotated reports as the training data for MIST. After testing MIST extensively on the training dataset using various parameterizations and per our consultations with its developers, we decided to run it with –4 bias, which greatly favors sensitivity over precision but not to the extent that the results become unreliable. We received great assistance from every member of the MIST developer team at every phase of our study.

In an earlier study,[14] we tested these systems using patient name information provided in HL7 segments. We plan to develop similar lookup mechanisms for eliciting and utilizing other patient identifiers. However, in the evaluation of this report, we did not use such PII that are available outside of medical reports.

Evaluation of differently tokenized results

Most de-identification systems come with their own tokenizers producing different sets of incompatible results. In order to compare the results and to report token misalignment errors, evaluators devised terminology such as colliding tokens, boundary detection failure and partially tagged tokens. For example, Deleger et al. reported that partially tagged PHIs due to boundary detection errors were 13% of all tagging errors.[6] Some researchers in the NLP community also use complex alignment schemas to remedy the problem.[24] When tokens produced by different systems do not match, the evaluation gets complicated and the differences between results become obscured. The situation gets complicated further as the number of systems to be compared increases. In the literature, we have not seen any proposed solution to the problem for robust evaluation of de-identification systems.

In this study, we align all outputs to be compared to the tokens of the gold standard. This method simplifies the evaluation without introducing any bias favoring one system over another: (1) We re-tokenize all outputs using the same tokenization scheme that the gold standard annotation has adopted. (2) When a token t in a system output does not correspond one-to-one to a gold standard token t_G, one of the following three scenarios is observed: (a) t may be a proper substring of t_G; (b) t_G may be a proper substring of t; or (c) t and t_G may overlap partially. After re-tokenization, the string of characters in t is distributed into a sequence of one or more tokens. We tag the resulting sequence of tokens with the original tag of t. (3) If t was tagged with a set of multiple tags originally, we apply them simultaneously to all tokens in the resulting sequence.

Mapping outputs to the gold standard is a type of normalization method. Normalization of the tokenized outputs allows us to evaluate every piece of output and compare it across all systems without compromise.

Nonparametric analytic methods

Confidence intervals are staples of biostatics where samples usually come from a well-known parametric distribution, where observations are random variables distributed independently and identically, which is not the case for words in our dataset.

To estimate confidence intervals (CIs) in this study, we adopted a non-parametric bootstrap method,[25] bias-corrected, accelerated (BC_a) percentile intervals as implemented in package *boot* in R. [26] Through a bootstrap resampling strategy, we could truly simulate our initial sampling method. For each bootstrap sample, we randomly selected a patient and then included all reports (hence all associated token sequences) of the patient into the sample. We repeated this process until reaching the same number of patients in our original test data.

We computed statistical significance for the scores where CIs were overlapping, based on Wilcoxon paired signed test with Pratt's adjustments,[27] using the package *coin* in R.[28] This method is more suitable than bootstrap based p value estimation because it can successfully take into account that two sets of compared results are paired datasets.

These methods can be used in a wide-range of computational linguistic studies and provide a strong analytic footing for comparisons of different study results. We previously used them to compare and analyze information extraction performances of various systems.[29]

Privacy risk analysis

Every identifier does not have the same value of information and is not equally revealing the identity of the patient. For example, revealing the patient's first name JOHN is not as significant as revealing the first name BARACK. While the re-identification risk calculation methods have been widely used in the anonymization and structured data de-identification context,[30] no other clinical text de-identification tool in the literature has been evaluated thoroughly involving the necessary risk analyses on the revealed or missed tokens. We introduced this notion in our earlier study on personal name de-identification, where we missed two name tokens.

The first missed token was a nickname of the patient's spouse. The recognizer missed the name, because we had not implemented a necessary mapping between nicknames and official names. Nicknames used as official names are rare in our name datasets; thus, the estimated probabilities for nicknames were unrealistic. We did the mapping manually, estimated the number of individuals in the U.S., who may be using that particular nickname and reported the result as our risk estimate.[14]

The second missed token was a severely distorted version of a patient's last name due to the transcriptionist's misspelling. There were two distinct misspellings on two different parts of the same token; thus, the name was unrecognizable. The word was used in English text frequently enough that the likelihood ratio score favored for the non-name hypothesis.

In order to perform the necessary risk analysis, we transformed the problem into the following question: If a person tries to figure out the actual name of the patient from the revealed misspelled version of the name, which names would s/he have to consider until arriving the actual name? Since the misspelled version required two edits, we searched all names within two-edit-distance from the misspelled version to the actual name, estimated the number of individuals who may have one of those names, and reported the result as our risk estimate.

Two out of 2388 patient name tokens missed is a necessary statistic but not sufficient; the research community and more importantly the public at large require more informative explanation for every potential breach of privacy. We believe it is the duty of de-identification system designers to establish some mechanisms to reliably calculate the involved privacy risk after such incidences occur. A comprehensive de-identification system should already have the necessary resources and tools to calculate the risk of re-identification.

Project Status

The NLM Scrubber is a stand-alone software system that accepts records in HL7 or free text format and outputs de-identified records. Version 1.0 of the system currently recognizes the following identifiers:

A. **Personal names and personal name initials**, corresponding to a subset of Privacy Rule identifiers in item 1 in Table 1
B. **Alphanumeric and telecom identifiers**, corresponding to identifiers in items 4–15 and 18 in Table 1
C. **Addresses**, corresponding to identifiers in item 2 in Table 1

D. **Dates** (incl. **ages**), corresponding to a subset of identifiers in item 3 in Table 1

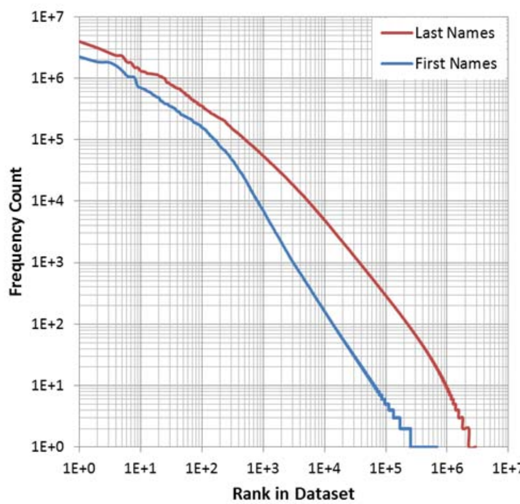
We have been working on employer name recognition, which corresponds to a subset of identifiers in item 1 in Table 1. We are also revising address and date recognition. Upon their completion and proper testing, the software package will be released to the public as free and open software. Identifiers in items 16 and 17 in Table 1 (i.e., biometric identifiers and images) are not available in narrative text reports; thus, they are not applicable to our design.

Evaluation

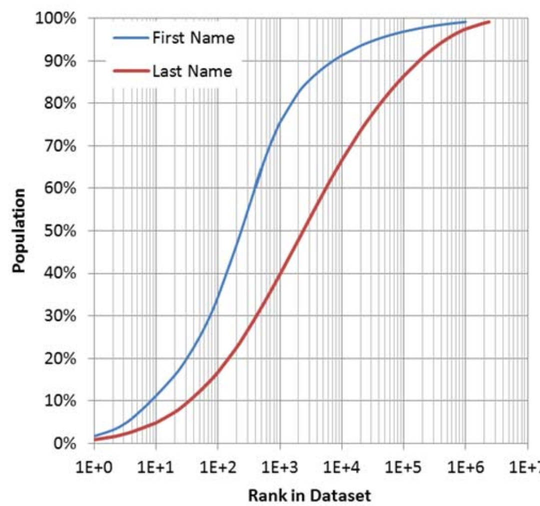
In this report, we chose to blend two aspects of our de-identification project. In addition to the NLM Scrubber, we also discussed a specific study of methods for de-identifying personal names, of which a substantial portion has been described by Kayaalp et al. elsewhere.[14] Below, we present the evaluation of NLM Scrubber in its current status.

Characteristics of the Datasets

The distributions of the first and last names in the name dataset are displayed in Figure 6. The distributions plotted in log-log coordinates in Figure 6(a) do not form straight lines as suggested by Zipf's Law, but for first names particularly, it could be estimated in two piecewise-linear functions at both sides of the inflection point at (300, 50 000). As seen in Figure 6(b), names of 50% U.S. population come from a set of less than 300 first names and less than 3000 last names.



(a)



(b)

Figure 6 Distributions of Personal Names in U.S. on (a) Log to Log Scale and (b) Log to Cumulative Linear Scale

In this report, we also tested the effectiveness of a simple false-positive filtering system. Despite its ad hoc nature and simplicity, the filtering system reduced 29% of all false positives patient names. Note however, these results need to be taken with a grain of salt and should give the reader only an optimistic perspective about the potential of the filtering, since the elements of the filter were compiled based on the results of our previous study.[14] The scientific value and the performance of the filtering system need to be validated on a new dataset.

Patient Names

The results of three de-identification systems, NLM Scrubber (NLM-S), MIT's de-identification system (MITdeid), and MITRE's de-identification system (MIST) are displayed in Table 5. The patient name de-identification performance of NLM-S and MITdeid was the same as in our earlier study, with the exception of decreased false positives. However, this new setup (in which we no longer labeled provider names as PHI) caused a severe performance problem for MIST despite the fact that we have not altered MIST parameters. In the earlier study, we reported 375 false negative patient names for MIST; whereas, in this study we see that that particular figure jumped to 615. We believe MIST's performance drop was due to the change in the training dataset. In this study, the training set contains 1050 patient name tokens. In the earlier study, we provided the same 1050 patient names in addition to 12 038 provider name tokens, all of which were lumped together within the set of personal name tokens. In other words, compared to the current study, there were 12.5 times as many data points to be trained on. These results indicates that the previous training set that included all types of personal names was much more beneficial for MIST than the current training set that strictly adheres to the gold standard.

Table 5 De-identification Performance Results of NLM Scrubber (NLM-S), MIT's de-identification system (MITdeid), and MIST: Bold fonts denote the best results among the three systems in columns Sensitivity, Specificity and Accuracy, which are also statistically significant if their confidence intervals are written in bold fonts.

Identifier	Gold	System	TP	FN	FP	TN	Sensitivity	Specificity	Accuracy
PatientName	2388	NLM-S	2386	2	24597	1117891	**0.999 (0.997,1.000)**	0.978 (0.978,0.979)	0.979 (0.978,0.979)
		MITdeid	2243	145	19482	1123006	0.939 (0.908,0.959)	0.983 (0.982,0.984)	0.983 (0.982,0.983)
		MIST	1773	615	3591	1138897	0.742 (0.685,0.786)	**0.997 (0.997,0.997)**	**0.996 (0.996,0.997)**
AlphaNumericId	4165	NLM-S	4163	2	8457	1132254	**1.000 (0.998,1.000)**	0.993 (0.992,0.993)	0.993 (0.992,0.993)
		MITdeid	1444	2721	1835	1138876	0.347 (0.333,0.359)	0.998 (0.998,0.998)	0.996 (0.996,0.996)
		MIST	4091	74	1804	1138907	0.982 (0.977,0.986)	**0.998 (0.998,0.999)**	**0.998 (0.998,0.999)**
Address	292	NLM-S	244	48	3466	1141118	0.836 (0.768,0.888)	0.997 (0.997,0.997)	0.997 (0.997,0.997)
		MITdeid	129	163	1428	1143156	0.442 (0.375,0.513)	0.999 (0.999,0.999)	0.999 (0.998,0.999)
		MIST	250	42	1174	1143410	**0.856** (0.791,0.905)	**0.999 (0.999,0.999)**	**0.999 (0.999,0.999)**
Date	29134	NLM-S	28823	311	730	1115012	0.989 (0.984,0.992)	**0.999 (0.999,0.999)**	**0.999 (0.999,0.999)**
		MITdeid	27595	1539	1094	1114648	0.947 (0.942,0.951)	0.999 (0.999,0.999)	0.998 (0.998,0.998)
		MIST	28906	228	2446	1113296	**0.992 (0.988,0.994)**	0.998 (0.997,0.998)	0.998 (0.997,0.998)

Employer	115	MIST	59	56	2750	1142011	0.513 (0.363,0.659)	0.998 (0.997,0.998)	0.998 (0.997,0.998)
		NLM-S	35820	274	33677	1075105	**0.992 (0.990,0.994)**	0.970 (0.969,0.971)	0.970 (0.969,0.971)
PHI	36094	MITdeid	31787	4307	23463	1085319	0.881 (0.875,0.886)	0.979 (0.978,0.98)	0.976 (0.975,0.977)
		MIST	35171	923	11673	1097109	0.974 (0.968,0.979)	**0.989 (0.989,0.99)**	**0.989 (0.988,0.971)**

All in all, NLM-S could catch all but 2 patient name tokens. The difference between sensitivity scores of NLM-S and MITdeid was statistically significant. Our privacy risk analysis indicates that the two false negatives had no significant impact on privacy of those two patients since the odds of re-identifying those individuals were less than 1 out of 150 000 in one case and less than 1 out of 200 000 in the other. MITdeid and MIST produced 72 and 307 times more false negative patient names, respectively.

Note that MIST's false positive rate (FPR) on patient names was impressively low, compared to 24 597 false positives (FPs) of NLM-S. On the other hand, 82% (20 149) FPs of NLM-S and 88% (17 133) FPs of MITdeid were provider names. In other words, a high FPR does not automatically indicate a loss of clinical information.

Likelihood Ratio Metric

The performance results of likelihood ratio metric are especially interesting. For each report, we extracted all patient name tokens and deleted the duplicates. After combining them from all reports, we observed that 98% patient names were labeled as personal name based on likelihood ratio statistics. Of the remaining 2%, 27% were nobiliary particles such as **van**, **di**, **de**, and **St**., and 20% were names that are frequently-occurring, regular English words such as **He**, **May** and **Day**.

Likelihood ratio metric performance was also quite good on non-PHI tokens—a non-PHI token is a token that is considered not PHI in the annotated (gold standard) dataset. More than 99.5% non-PHI tokens were labeled correctly as non-names based on likelihood ratio statistics. Of the remaining tokens, 55% were starting with lower case letters. Thus, 99.8% of non-PHI tokens were preserved by our likelihood ratio metric and our initCap rule. Recall that initCap rule filters in only capitalized initial letter tokens as names.

Alphanumeric Identifiers

In alphanumeric identifiers, NLM-S performance was clearly superior to others. It missed only two tokens, one of which was "406," a 3-digit area code of a telephone number, which should be considered non-PII since the area it covers is the entire state of Montana.[31] The other missed token was a protocol number, which is considered a low risk to privacy as the necessary information to re-identify the patient is not publicly available and such information is usually given to the patient's health care providers only.[32] MITdeid did not produce a viable alphanumeric de-identification on this dataset.

Addresses and Dates

In both addresses and dates, MIST results yield the best sensitivity and specificity scores, but on addresses, the sensitivity score difference between NLM-S and MIST was not statistically significant. After reviewing the false negative cases of NLM-S, we observed that most of the NLM-S's "missed address tokens" were actually non-PHI tokens such as geographical direction (e.g., Northern), state name abbreviation (e.g., VA), large city names in other countries (e.g., Beijing) and country names (e.g., England). None of the missed address tokens was revealing a street address, but three of the revealed address tokens may cause some privacy concerns. They were Falls Church (Falls Church, VA: pop. 12 751) and Takoma (Takoma Park, MD: pop. 17 021). Note that we are not revealing any PHI here; demographic information was not connected to any health information of an individual.

In dates, NLM-S was clearly superior in terms of specificity and accuracy as the differences of the corresponding NLM-S and MIST scores were statistically significant, which however have far less importance than the sensitivity score and NLM-S requires further sensitivity improvement on dates. None of the revealed dates was tagged as PII-Age (i.e., age > 89) in the gold standard.

Although trailing behind the other two systems, MITdeid showed strong sensitivity performance on dates (0.947), but not on addresses (0.442).

Employer Names

Since neither NLM-S nor MITdeid implemented employer name recognition, only MIST's results were tabulated in Table 5; however, MIST's sensitivity (0.513) was below an acceptable range.

Overall Performance

It is not uncommon that a system tags a PHI token (e.g., a date) with a wrong PII label (e.g., an alphanumeric identifier). In such cases, there is neither a leakage of PHI nor a loss of clinical information. The PHI row in Table 5 indicates that there were a total of 36 094 PHI tokens, of which NLM-S missed only 274, MIST 923 (3.4 times as many), and MITdeid 4 307 (15.7 times as many). NLM-S was clearly superior in overall sensitivity and MIST was clearly superior in overall specificity and accuracy.

Note that in Table 5 we reported NLM-S's PHI specificity as 0.97. Although accurate, it could be misconstrued easily by readers who do not pay attention to the false positive (FP) count, 33 677, of which 20 149 were provider names, which do not constitute clinical information. After excluding those FP provider name counts, we end up with an FP count of 13 528 and a healthy specificity score of 0.988, which is comparable to the best specificity score of 0.989.

The decomposition of the revealed PHI tokens by PII types is displayed in Table 6. Note that the superiority of MIST that we observed in Table 5 was totally washed away in Table 6, where the best performer became NLM-S in all identifier recognition tasks that we implemented in NLM-S. We will complete NLM-S by implementing employer name recognition. As seen in these results, the most problematic identifiers for NLM-S were dates and addresses. Although NLM-S outperformed the other two systems in dates and addresses, we still need to revise and re-implement those recognizers in order to make NLM-S a robust de-identification system.

Table 6 Decompositions of Revealed PHI tokens by System

Tag	PII Type	Gold	NLM-S	MIST	MITdeid
Name	**Patient**	2387	2	592	136
	PNInit	1	0	1	1
	Employer	115	92	54	101
AlphaNumericId	**AlphaNumericId**	3502	0	24	1885
	Protocol-Id	660	1	3	659
	Telecom	3	1	0	0

Address	**Address**	292	29	38	63
Date	**Date**	29124	149	204	1457
	Age 90+	10	0	7	5
All		**36094**	**274**	**923**	**4307**

Discussion

As seen in results, NLM-S incurred substantial number of false positives in order to catch the maximum number of identifiers. Our primary goal and our main criterion for success are to eliminate all PHI tokens when possible. For keeping the trust of the U.S. Public to the research community, we have to continue working on improving the sensitivity of NLM-S even if it costs us more false positives to achieve that. On the other hand, we are also cognizant to the needs of the research community and have to pay great attention to false positive rates and to the effective conservation of clinical information in the upcoming versions of NLM-S.

In our study data, NLM-S has recognized more PHI tokens than MIST and MITdeid have, which are the only freely available, general-purpose clinical text de-identification systems at the time of this publication. Our risk analysis indicates that the revealed tokens would not cause any substantial risk to the patient privacy. Only three instances of address identifiers revealed the home city of three distinct patients, where the population sizes were less than 20 000 but greater than 12 500. Population size 20 000 was devised by the Privacy Rule as a threshold for further censoring zip codes (see Table 1).

MIST was clearly the second best performing system of this study. Due to their underlying methodological power, probabilistic machine learning systems do very well in this domain. Given that we devised our system based on the characteristics of the clinical corpus in our hand, we should not be surprised if MIST outperforms NLM-S in another clinical dataset with different characteristics.

As we indicated in one of our earlier studies,[29] probabilistic machine learning and symbolic linguistic methods are not an either-or proposition, a good NLP system should incorporate methods of both paradigms and reap the benefits of both worlds. The success of our likelihood ratio metric is a good example for that. We plan to develop a robust machine learning component to our scrubber so that it could perform well on a variety of reports from different origins.

Distribution and Use Case Scenarios

A de-identification software system may be used in various ways, which may influence the development and determine the minimum level of sensitivity and specificity expected from the system. For example, if the de-identification software is planned to be used to aid a human expert for semi-automatic de-identification or clinical report annotation (e.g., to create a training dataset), depending on the expert's preferences, the false negative or false positive rate might not be as an important concern as it would be in the fully automated case.

There are three distinct potential distribution/use case scenarios for a de-identification system: (1) distribution of the system as a stand-alone application, (2) de-identification as an online service, and (3) providing de-identified data to researchers. Since they are orthogonal to each other, any combination of these three scenarios is also conceivable.

The usual scenario is to distribute the de-identification software as a stand-alone application. This is also our current plan. We further plan to release the source code to the public after necessary tests and evaluation. In this scenario, we may provide software patches and improvements in the subsequent releases.

To the best of our knowledge, the second distribution scenario, de-identification as an online service, has not been made available by any software developing institution or company. In this scenario, instead of releasing software patches and updates, we can monitor its performance actively, correct the errors immediately, and improve its performance continuously.

The third scenario, where not the software but the de-identified data is provided, is a rare one. Its only actual example that we are aware of is the MIMIC (Multi-parameter Intelligent Monitoring for Intensive Care) database.[33] MIMIC II Waveform Database is distributed freely. Researchers can also access the clinical data in the MIMIC II Clinical Database if their application is accepted by the provider of the data after they agree the terms and conditions set forth by the provider.[34]

All scenarios have pros and cons. In the first scenario, it is hard to predict the de-identification performance of the system in an arbitrary setting. The variance of de-identification performance could be held under control, if the system is used by major clinical centers in the U.S., given our familiarity with the settings and the clinical culture in the U.S. If however the system is used in a niche clinical center or in another English speaking country, e.g. South Africa, it would be hard to predict the performance and the variance on different types of clinical reports.

An advantage of the first scenario is that the system is easier to build and release. It is possible that others may analyze the code, discover its weaknesses and suggest some improvements. In that scenario, we could reap the benefit of the open-source nature of the process. On the other hand, it is also possible that it may be used by parties who are vaguely familiar with the application and with its settings and may require support beyond our resources.

The second scenario, providing de-identification as an online service, is a more expensive undertaking as it requires significant infrastructure and support personnel. However, there are certain benefits that we cannot attain in the first scenario: (1) we would be able to conduct continuous quality checks, tune our application given the continuous stream of data coming from various sources, and build a set of machine learning models progressively; (2) to further improve the performance, we can use proprietary data and tools that we cannot distribute freely; (3) by signing into a data use agreement, we can put in place certain mandatory reporting mechanisms that alert us about de-identification failures or other improper system behavior, which in turn would help us eliminate the weaknesses of our system and tune it in a timely manner.

Another advantage of the second scenario is that it would require no infrastructure resource for the user. Users with little or no familiarity with the software and the operating system environment may be able to use the system and easily create their de-identified dataset. This option would be very beneficial to researchers in local hospitals and other small institutions; therefore, in this scenario a wider biomedical research community can access de-identified health information.

In the third scenario, the ease of access to de-identified data would be increased further since the user no longer needs to tackle with the de-identification process and can access to the de-identified health information directly.

In this case, the user space would include not only the clinical community but also other research communities such as medical informatics community and computer scientists. Since the data can be accessed much more widely than the other two scenarios, the de-identification of the data and data use agreement must be most stringent.

Scientific Contributions

1. **Likelihood ratio metric**: We developed an innovative method to compute likelihood ratio of name to non-name of a given word. Toward this end, we used disparate sources of datasets: a name frequency data of a large group of Social Security applicants in the U.S. and two English corpora, Wikipedia and Medline abstracts of core clinical journals.
2. **Method for Sampling Clinical Reports**: We devised a new sampling method that preserves the randomness of the patients and eliminate duplications of their reports. None of the de-identification studies in the literature and no other clinical-corpus-based studies that we know of have analyzed their data in such close scrutiny as we did or proposed a comparable sampling method. The sampling method of our study may guide the future research on the big clinical text data.
3. **Nonparametric analytic methods**: In corpus linguistics research, where the main unit of data is word, results are almost always provided in absolute values, because, due to the intrinsic dependencies among words, one could not assume that words are random variables distributed independently and identically. Simulating our sampling method of clinical reports (hence sequences of words), we devised a bootstrap sampling schema, which provides reliable confidence intervals. We also proposed to use another nonparametric method, Wilcoxon paired sign test using Pratt's adjustments, which provides reliable p values for our results and can do the same for the results of similar NLP systems.
4. **Normalization of differently tokenized results**: De-identification systems usually come with their own tokenizers; thus, the results of different de-identification systems usually are misaligned. No de-identification research article has ever addressed this issue. We proposed a simple unbiased method to align all de-identification outputs to the annotated (gold standard).
5. **Privacy risk analysis**: De-identification system performance statistics provide an overall impression how the system behaves, but they fall short to explain about how much privacy risk each revealed PHI token introduces. In this study, we proposed solutions to estimate such risks for revealed name, address tokens, and, in special circumstances, for partial revelation of alphanumeric identifiers.

Summary and Future Plan

We reported four major and one minor distinct scientific contributions of the project (see *Scientific Contributions* for a summary). One of those contributions is calculating the risk of re-identification of different revealed tokens. Revealing PHI tokens should not be treated as statistics only. The research community and, more importantly, the public at large require more informative descriptions for every potential breach of privacy. We believe it is the duty of the de-identification system designers to establish some mechanisms to reliably calculate the involved privacy risk when such incidences occur.

Although NLM-S overall performance was superior to the other two de-identification systems, there is still room for improvement especially in address and date recognition. Note however, both types of information (except street addresses) can be revealed to researchers if they sign into a data use agreement with the provider of the data; hence, they are not as critical as patient names and alphanumeric identifiers.[4] We will implement nickname and employer name recognition modules before releasing the software package to the public.

The legacy code that NLM-S partially relies on contains a large amount of unnecessary code that is difficult to maintain. We plan to eliminate the legacy code as soon as we can and to cut the dependencies of the working code on it.

As our studies indicate, MIST performance clearly depends on the size and quality of the training set. Machine learning systems like MIST are more tolerant and adaptive to the change of the corpus. Although in our studies NLM-S outperformed MIST in protecting PHI, we cannot guarantee the same if the composition and the characteristics of the clinical corpus are changed significantly.

We plan to maintain NLM-S's performance across a wide range of clinical documents by incorporating probabilistic learning modules. We also plan to test NLM-S on different report types as well on clinical text reports of other institutions outside of NIH.

Finally, in our annotation schema, we have not specified dates in more details such as noting the date of birth, hospital admission and discharge dates of the patient. Such dates have greater significance than common dates such as the date of the report or the date of the test, because they might be found in financial records of the patient and could be linked to the patient. Therefore, revealing such dates poses a greater risk to privacy. We plan to annotate our dataset accordingly in order to evaluate associated privacy risks more reliably.

Project Schedule and Resources

Our team composition is listed in Table 7.

Table 7 Project Team Composition

Staff Years	Name	Main Role
0.8	Mehmet Kayaalp	Lead Investigator
1.0	Zeyno Dodd	Programmer
0.3	Allen Browne	Linguist/Annotator
0.1	Pamela Sagan	Nurse Annotator
0.1	Clement McDonald	Supervisor of Record
2.3	**Total Human Resources**	

Our current schedule is provided in Table 8.

Table 8 Development and Release Schedule of the Project

	Task	Date
1	BTRIS Pilot: NIH Clinical Center pilot deployment	October 2013
2	NLM Scrubber version 1.0 release with employer name de-identification	February 2014
3	NLM Scrubber version 2.0 release without legacy code	February 2015
4	NLM Scrubber version 3.0 release with adaptive learning components	February 2016

Collaborations

During the lifetime of the project, we have collaborated with the following groups:

1. Lynette Hirschman's group at MITRE in Boston, developers of MIST[12]
2. Jeff Friedlin from Regenstrief Institute[35]
3. Social Security Administration, from where we obtained name frequency data of social security applicants
4. NIH Clinical Center, from where we obtained the study data
5. BTRIS at NIH: We are working on deploying our first test pilot

Contributorship Statement

Clement McDonald initiated the project, set the design principle (i.e., a dictionary based system), and supervised the project. Mehmet Kayaalp designed and implemented personal name and alphanumeric identifier recognition components. He also identified and acquired all related datasets and designed their integrations with contributions of Selcuk Ozturk. Mehmet Kayaalp also designed the study, the evaluation methods, and authored this report. Yanna Kang and Zeyno Dodd contributed to the final coding of Bootstrap and Wilcoxon paired signed rank test in R language. Clement McDonald set the content of false positive name filtering, which was implemented by Zeyno Dodd. Allen Browne and Guy Divita designed dTagger and Guy Divita implemented it with contributions of Russell Loane. Guy Divita also implemented the date and address recognition modules. Zeyno Dodd implemented the age recognition module, ran all experiments, and collected their results. Mehmet Kayaalp set the requirements specification and initial design principles of the annotation tool VTT. Allen Browne, Guy Divita and Chris Lu contributed to the design of VTT and Chris Lu implemented the code. Allen Browne and Pam Sagan annotated the clinical corpus.

Acknowledgements

We are grateful to Bert M. Kestenbaum of the Office of the Chief Actuary at the Social Security Administration, who provided us frequencies of names of applicants for a social security number, and Dr. Jon McKeeby, CIO, Clinical Center at NIH and his staff for their help in obtaining and interpreting the clinical data. We acknowledge contributions of Guy Divita to the design and coding of dTagger, Address and Date recognition. We also acknowledge contributions of Dr. Chris Lu to the design and coding of VTT, which has been central to our annotation efforts. We also acknowledge contributions of Dr. Yanna Kang to the development of the R codes of this study. We also acknowledge various contributions of Selcuk Ozturk and Shuang Cai. We also thank Drs. Lynette Hirschman, Samuel Bayer and Ben Wellner as well as John Aberdeen of MITRE, for their generous help and offers to test MIST on our study data.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of Interest Disclosure

The author of this report, who is also the lead investigator of the project, receives royalties from University of Pittsburgh for his contribution to a de-identification project. The resulting product was acquired by a third party, which today is known as the De-ID Data Corp. Prior joining to the project, he had disclosed the information to NLM's Ethics Office, where his case was reviewed and his appointment was approved.

Glossary

Legal definitions provided here are based on the text of 45 CFR Subtitle A § 160.103 (10-1-10 Edition).

alphanumeric string	string of characters that contains one or more digits and may also contain other characters
covered entity	(1) A health plan. (2) A health care clearinghouse. (3) A health care provider who transmits any health information in electronic form in connection with a transaction covered by 45 CFR § 160.103 (see below for full description of *transaction*).
de-identification	removal of PII that is part of PHI from data (see *personally identifiable information* and *protected health information*)
finite state automaton (FSA)	state machine represented in a directed graph where states are represented in vertices, transitions in directed arcs, and inputs causing the transitions in labels on the arcs. An input sequence is accepted by an FSA if the first element of the sequence causes a transition from the start state to another state of FSA and its last element reaches to the end state of FSA.
health care clearinghouse	a public or private entity, including a billing service, repricing company, community health management information system or community health information system, and ''value-added'' networks and switches, that does either of the following functions: (1) Processes or facilitates the processing of health information received from another entity in a nonstandard format or containing nonstandard data content into standard data elements or a standard transaction. (2) Receives a standard transaction from another entity and processes or facilitates the processing of health information into nonstandard format or nonstandard data content for the receiving entity.
health information	any information, whether oral or recorded in any form or medium, that: (1) Is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and (2) Relates to the past, present, or future physical or mental health or condition of an individual; the provision of

	health care to an individual; or the past, present, or future payment for the provision of health care to an individual.
individual	the person who is the subject of protected health information
individually identifiable health information	information that is a subset of health information, including demographic information collected from an individual, and: (1) Is created or received by a health care provider, health plan, employer, or health care clearinghouse; and (2) Relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual; and (i) That identifies the individual; or (ii) With respect to which there is a reasonable basis to believe the information can be used to identify the individual.
personally identifiable information (PII)	information that identifies a person directly (e.g., social security number, personal name) or indirectly (e.g., home address or other affiliations to a small group of people); synonym: individually identifiable information
PHI token	an alphanumeric token or word that contains PII as part of PHI. Identifiers that are not PHI (e.g., physician license number) are not labeled as PHI tokens; complement: non-PHI token (see *token*)
protected health information (PHI)	individually identifiable health information: (1) Except as provided in paragraph (2) of this definition, that is: (i) Transmitted by electronic media; (ii) Maintained in electronic media; or (iii) Transmitted or maintained in any other form or medium. (2) Protected health information excludes individually identifiable health information in: (i) Education records covered by the Family Educational Rights and Privacy Act, as amended, 20 U.S.C. 1232g; (ii) Records described at 20 U.S.C. 1232g(a)(4)(B)(iv); and (iii) Employment records held by a covered entity in its role as employer.
token	a sequence of characters, such as a word, a number, punctuations, or any combination of these, which may serve as a lexical unit in the analysis
transaction	transmission of information between two parties to carry out financial or administrative activities related to health care. It includes the following types of information transmissions: (1) Health care claims or equivalent encounter information. (2) Health care payment and remittance advice. (3) Coordination of benefits. (4) Health care claim status. (5) Enrollment and disenrollment in a health plan. (6) Eligibility for a health plan. (7) Health plan premium payments.

(8) Referral certification and authorization.
(9) First report of injury.
(10) Health claims attachments.
(11) Other transactions that the Secretary (of Health and Human Services or any other officer or employee of HHS to whom the authority involved has been delegated) may prescribe by regulation.

VTT Visual Tagging Tool, a GUI based annotation tool, which is part of NLM Scrubber v.1.0. It also implies the format of the text structure presented in Figures 1–3.

References

1. U.S. Census Bureau. ZIP Code Tabulation Areas, 2010.

2. Department of Health and Human Services. Public Welfare; Administrative Data Standards and Related Requirements; General Administrative Requirements; General Provisions; Definitions. 45 CFR § 160.103.

3. McCallister E, Grance T, Scarfone K. Guide to Protecting the Confidentialiyt of Personally Identifiable Information (PII). Recommendations of the National Institute of Standards and Technology. U.S. Department of Commerce, NIST, 2010.

4. Department of Health and Human Services. Public Welfare; Administrative Data Standards and Related Requirements; Security and Privacy; Privacy of Individually Identifiable Health Information; Other Requirements Relating to Uses and Disclosures of Protected Health Information. 45 CFR § 164.514.

5. Meystre S, Friedlin F, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology* 2010;10(1):70.

6. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *J Am Med Inform Assoc* 2013;20(1):84-94.

7. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *J Am Med Inform Assn* 2013;20(1):77-83.

8. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Med Inform Decis Mak* 2006;6:12.

9. Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. *Arch Pathol Lab Med* 2003;127(6):680-686.

10. Gardner J, Xiong L. HIDE: An Integrated System for Health Information De-identification. *Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems* 2008:254-259.

11. Neamatullah I. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008;8:32.

12. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly Retargetable Approaches to De-identification in Medical Records. *J Am Med Inform Assn* 2007;14(5):564–573.

13. Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. *J Am Med Inform Assn* 2007;14(5):550-563.

14. Kayaalp M, Browne AC, Callaghan FM, Dodd ZA, Divita G, Ozturk S, et al. The Pattern of Name Tokens in Narrative Clinical Text and a Comparison of Five Systems for Redacting them. *J Am Med Inform Assn* 2013.

15. National Library of Medicine. Visual Tagging Tool, 2010. URL: http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/web/index.html. Accessed in 8/20/2013

16. Kayaalp M. Seperation of Data, Interpreters and Likelihood. Technical Report LHNCBC-TR-2007-001: The Lister Hill National Center for Biomedical Communications, 2007. URL: http://lhncbc.nlm.nih.gov/files/archive/tr2007001.pdf. Accessed in 9/2/2013.

17. Social Security Administration's Death Master File: National Technical Information Service, U.S. Department of Commerce. URL: http://www.ntis.gov/products/ssa-dmf.aspx. Accessed in 10/11/2011.

18. MEDLINE®/PubMed® Resources Guide: U.S. National Library of Medicine, National Institutes of Health. URL: http://www.nlm.nih.gov/bsd/pmresources.html. Accessed in 10/11/2011.

19. Wikipedia: Database download: Wikipedia Foundation, Inc. URL: http://en.wikipedia.org/wiki/Wikipedia:Database_download. Accessed in 10/11/2011..

20. Abridged Index Medicus (AIM or "Core Clinical") Journal Titles: U.S. National Library of Medicine, National Instititutes of Health. URL: http://www.nlm.nih.gov/bsd/aim.html. Accessed in 10/11/2011.

21. Bernardo JM, Smith AFM. *Bayesian Theory*: John Wiley & Sons, 1994.

22. Jaynes ET. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics* 1968;sec-4(3):227-41.

23. Hopcroft JE, Ullman JD. *Introduction to Automata Theory, Languages, and Computation*: Addison-Wesley, 1979.

24. NTT System Description for the WMT 2006 Shared Task. *Workshop on Statistical Machine Translation*; 2006; New York, NY. Association for Computational Linguistics.

25. Efron B. Better Bootstrap Confidence Interval. *Journal of the American Statistical Association* 1987;82(397):171-185.

26. Davison AC, Hinkley DV. *Bootstrap Methods and their Application*: Cambridge University Press, 1997.

27. Pratt JW. Remarks on Zero and Ties in the Wilcoxon Signed Rank Procedures. *Journal of the American Statistical Association* 1959;54(287):655-667.

28. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. Implementing a Class of Permutation Test: The coin Package. *Journal of Statistical Software* 2008;28(8):1–23.

29. Kang YS, Kayaalp M. Extracting Laboratory Test Information from Biomedical Text. *Journal of Pathology Informatics* 2013;4(1):23-35. URL: http://www.jpathinformatics.org/ text.asp?2013/4/1/23/117450. Accessed in 9/3/2013.

30. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assn* 2010;17(2):169-177

31. Wikipedia. Area code 406: Wikipedia, 2013. URL: http://en.wikipedia.org/wiki/Area_ code_406. Accessed in 8/20/2013.

32. Office of Civil Rights. Guidance Regarding Methods for De-idnetification of Protected Health Information in Accordance with Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, 2012.

33. Moody GB, Mark RG. A Database to Support Development and Evaluation of Intelligent Intensive Care Monitoring *Computers in Cardiology* 1996;23:657-660.

34. PhysioNet. MIMIC II: Getting Access.URL: http://physionet.org/mimic2/mimic2_access .shtml. Accessed in 9/3/2013.

35. Friedlin F, McDonald C. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc* 2008;15(5):601 -610.